ANALYTIC TOOLS FOR LINGUISTS

Oana David
odavid@ucmerced.edu
Class lecture COGS 5
Thursday, April 6th, 2017

There is an exam on Tuesday 4/11

There is a study guide on Cat Courses!

What is linguistic 'data'?

Lexicon - words

Grammar – the structure of sentences

Naturalistic data – conversations, texts

Spoken: conversational exchanges, TV newscasts, interviews

Written: newspapers, books, Twitter, blogs

Approaches and Tools

- Corpus Linguistics
- Discourse Analysis
- Semantic Analysis

Corpus Linguistics

- What is a '**corpus**'? –
  - A digital collection of texts
  - Gathered for a particular purpose
  - That is searchable
  - Annotated for – date/period, author, genre, register, language

  - **Balanced corpus**: a collection of texts from all types of sources, to give a representative sample of the language as a whole.

  - **Specialized corpus**: a corpus put together for a specific purpose (e.g. Corpus of Old English, TV News Archive)

Corpus Linguistics

Text → searchable text

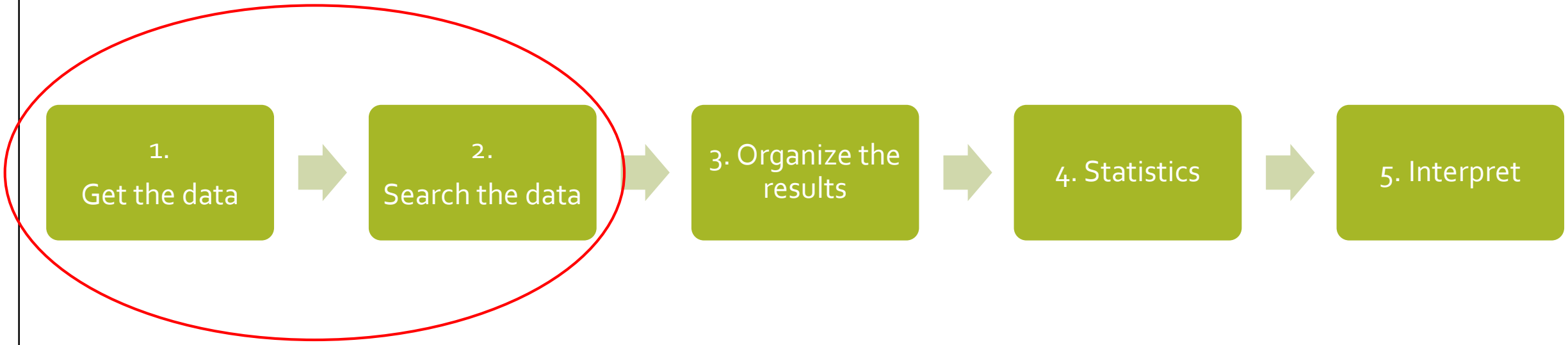
Tokenizing: breaking up strings into words, phrases, spaces, symbols, etc.

Lemmatizing: turning words into lemmas (e.g. ‘attack’ and ‘attacks’ are the same word)

Part of speech tagging: nouns, verbs, adj. (e.g. ‘attack’ as a verb and as a noun)

Parsing: situating words in their sentential context – subjects, objects, etc.

How do you 'do' corpus linguistics?



Step 0: What is the research question? What are you looking for, why, and what do you expect to find?

The BYU Corpora

Mark Davies

Brigham Young University

http://corpus.byu.edu/

English	# words	language/dialect	time period	compare
NOW Corpus	4.2 **billion+**	20 countries / Web	2010-**yesterday**	
Global Web-Based English (GloWbE)	1.9 **billion**	20 countries / Web	2012-13	
Wikipedia Corpus	1.9 **billion**	English	-2014	Info
Hansard Corpus (British Parliament)	1.6 **billion**	British	1803-2005	Info
Corpus of Contemporary American English (COCA)	**520** million	American	1990-2015	* * * * *
Corpus of Historical American English (COHA)	400 million	American	1810-2009	* *
Corpus of US Supreme Court Opinions NEW	130 million	American	1790s-present	
TIME Magazine Corpus	100 million	American	1923-2006	
Corpus of American Soap Operas	100 million	American	2001-2012	*
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993	* *
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s	
CORE Corpus	50 million	Web registers	-2014	
Other languages				
Corpus del Español (see also...)	2.1 **billion**	Spanish	1200s-1900s	*
Corpus do Português (see also...)	1.1 billion	Portuguese	1300s-1900s	
N-grams				
Google Books: American English	155 **billion**	American	1500s-2000s	*
Google Books: British English	34 billion	British	1500s-2000s	
Google Books: Spanish	45 billion	Spanish	1500s-2000s	

COCA: Corpus of Contemporary English

Search: “attack” -- Concordance

CLICK FOR MORE CONTEXT							[?]
1	2015	ACAD	JAdolAdultLiteracy	A	B	C	I see the Camras teachers, what I see is a bunch of teachers who **attack** a problem together with the mindset that a solution exists, you just have to
2	2015	ACAD	GeneralMusicToday	A	B	C	modes. In this practice, the performer bends a sustained pitch after the initial **attack** as a way to develop the mood of the raga. Teachers can determine which
3	2015	ACAD	ReadingImprovement	A	B	C	Direct instruction begins with simple sound-symbol relationships and logically progresses to phonetic rules and word **attack** strategies. # There are numerous
4	2015	ACAD	ReadingImprovement	A	B	C	2639 Students beyond the primary grades who struggle with reading lack basic word **attack** ability and are thus unable to read new words with accuracy. How
5	2015	ACAD	ReadingImprovement	A	B	C	with accuracy. However, there has been little research that reveals the specific word **attack** errors students exhibit. In this study, 123 struggling adolescent rea
6	2015	ACAD	ReadingImprovement	A	B	C	Consortium on Reading Excellence, Inc. (CORE) Phonics Survey to determine word **attack** knowledge for both real and pseudo one syllable and multi-syllabic w
7	2015	ACAD	ReadingImprovement	A	B	C	to apply these basic reading or decoding skills is sometimes referred to as having word **attack** skills. Ideally, it is an ability that should be learned in the primar
8	2015	ACAD	ReadingImprovement	A	B	C	Stanovich, 1984,1986). # Although it is ideal for students to acquire word **attack** skills in the primary grades, some students struggle to gain them and are pron
9	2015	ACAD	ReadingImprovement	A	B	C	such words as assuaged, foray, impudent and taciturn. Students without foundational word **attack** skills are faced with an even greater challenge reading thes
10	2015	ACAD	ReadingImprovement	A	B	C	struggling readers confirms what teachers have suspected: many of them do lack basic word **attack** skills even though there is a persistent myth that the early
11	2015	ACAD	ReadingImprovement	A	B	C	, but many struggling readers also demonstrated significant deficits at the word level (word **attack**, decoding, word recognition, and rate). # Another comparis
12	2015	ACAD	ReadingImprovement	A	B	C	skills. For example, with regard to the ability to decode words (word **attack**), they do not show whether students struggle more often with decoding unknown
13	2015	ACAD	ReadingImprovement	A	B	C	for them. This study was conducted to provide some information about the specific word **attack** skills with which struggling adolescent readers have the most
14	2015	ACAD	ReadingImprovement	A	B	C	this study's research questions were: 1) How did the students' word **attack** ability differ across the various subscales of the Consortium on Reading Excellence
15	2015	ACAD	ReadingImprovement	A	B	C	their ability to read pseudo words and what does this difference indicate about their word **attack** skills? 3) Which student characteristics accounted for differe
16	2015	ACAD	ReadingImprovement	A	B	C	word attack skills? 3) Which student characteristics accounted for differences in their word **attack** skills? # Method # Participants # In this study, 123 middle sc
17	2015	ACAD	ReadingImprovement	A	B	C	1990). # The one syllable reading/decoding section of the test includes five word **attack** subscales: 1) short vowels in Consonant Vowel Consonant (CVC) words

Search: “attack” -- Collocations

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...] COMPARE

		CONTEXT	FREQ		ALL	%	MI	
1		HEART	5792		100253	5.78	6.22	
2		UNDER	2073		239340	0.87	3.48	
3		TERRORIST	1549		16483	9.40	6.92	
4		DIED	637		68965	0.92	3.57	
5		IRAQ	623		78493	0.79	3.35	
6		ATTACK	605		51819	1.17	3.91	
7		KILLED	513		62275	0.82	3.41	
8		STROKE	466		11614	4.01	5.69	
9		NUCLEAR	460		46329	0.99	3.68	
10		LAUNCHED	368		16204	2.27	4.87	
11		ISRAEL	350		46772	0.75	3.27	
12		SUFFERED	340		18130	1.88	4.60	
13		HARBOR	336		11752	2.86	5.20	
14		PEARL	328		8008	4.10	5.72	
15		MISSILE	298		9865	3.02	5.28	
16		TERROR	281		15874	1.77	4.51	
17		ADS	279		13693	2.04	4.72	
18		LAUNCH	278		14749	1.88	4.60	
19		SURPRISE	277		30805	0.90	3.53	

Search: “attack”

SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
SPOKEN	16,888	109.4	154.38	
FICTION	5,963	104.9	56.84	
MAGAZINE	10,524	110.1	95.58	
NEWSPAPER	11,947	106.0	112.75	
ACADEMIC	6,439	103.4	62.26	
1990-1994	9,965	104.0	95.82	
1995-1999	7,830	103.4	75.69	
2000-2004	12,388	102.9	120.34	
2005-2009	9,335	102.0	91.48	
2010-2015	12,243	121.6	100.71	
TOTAL	**103,522**			SEE ALL TOKENS

Search: “attack” and “fight”

WORD 1 (W1): **ATTACK** (0.90)

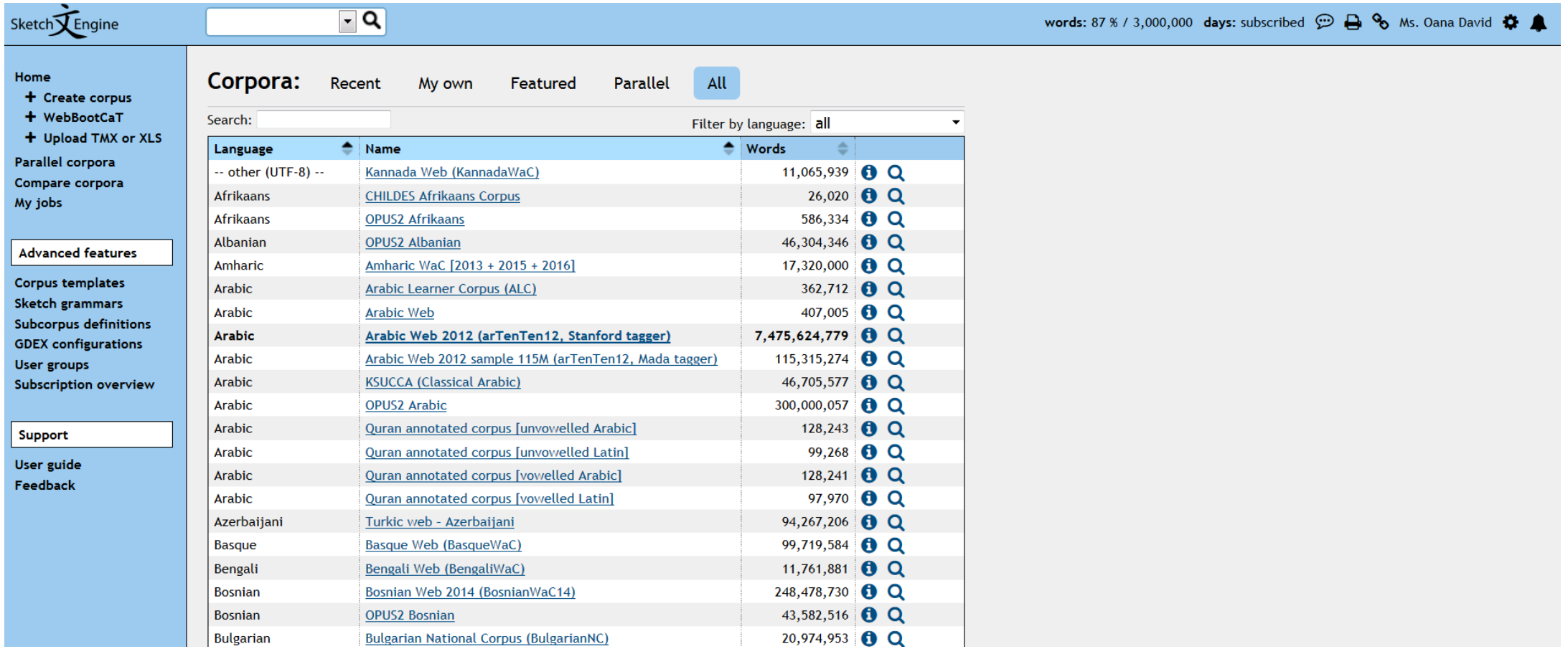
	WORD	W1	W2	W1/W2	SCORE
1	HARBOR	336	0	672.0	742.6
2	PEARL	328	0	656.0	724.9
3	TARGETS	170	0	340.0	375.7
4	BENGHAZI	107	0	214.0	236.5
5	MORTAR	99	0	198.0	218.8
6	EMBASSY	85	0	170.0	187.9
7	SNEAK	82	0	164.0	181.2
8	FRONTAL	78	0	156.0	172.4
9	MISSILE	298	2	149.0	164.7
10	DEADLIEST	72	0	144.0	159.1
11	ADS	279	2	139.5	154.2
12	CONVOY	63	0	126.0	139.2
13	USS	63	0	126.0	139.2
14	UNPROVOKED	59	0	118.0	130.4
15	STROKE	466	4	116.5	128.7
16	SCUD	56	0	112.0	123.8
17	GRENADE	55	0	110.0	121.6
18	KERRIGAN	53	0	106.0	117.1

WORD 2 (W2): **FIGHT** (1.11)

	WORD	W2	W1	W2/W1	SCORE
1	CUSTODY	113	0	226.0	204.5
2	SCENES	85	0	170.0	153.8
3	TYSON	84	0	168.0	152.0
4	HEAVYWEIGHT	82	0	164.0	148.4
5	TOOTH	79	0	158.0	143.0
6	FIST	74	0	148.0	133.9
7	EQUALITY	68	0	136.0	123.1
8	PROXY	68	0	136.0	123.1
9	NAIL	65	0	130.0	117.6
10	CHAMPIONSHIP	61	0	122.0	110.4
11	SNOWBALL	61	0	122.0	110.4
12	FIGHTS	45	0	90.0	81.4
13	COUPLES	43	0	86.0	77.8
14	FOREVER	43	0	86.0	77.8
15	COMMUNISM	42	0	84.0	76.0
16	VOW	37	0	74.0	67.0
17	BATTLES	216	3	72.0	65.2
18	BOXING	34	0	68.0	61.5

Sketch Engine

https://the.sketchengine.co.uk/



offering their lands as "launching pads" for a military	attack	. The spies apparentl;y believe that such
k. The spies apparentl;y believe that such an	attack	would be led by Britain - an indication of paranoia
and was found in the couple's house. He too had been	attacked	with ferocity, but he was still alive, and is now un
He has told his doctors that their home was	attacked	by ten men last week. He knew none of them by s
reason with the men but were ignored. "They began	attacking	them with pick handles and sticks, and both collap
" There seems to be no clear motive for the	attack	. Local villagers described the Austens as "a very u
/p> The angry soldiers moved into the streets and	attacked	several illegal dealers in foreign exchange who ope
him. What an irresponsible prat! He was subsequently	attacked	by just made about everyone in the office and I th
ereon ." In exchange, the Indians agreed to stop their	attacks	and raids. Ten chiefs endorsed the treaty and mar
nites. Thus, a plan was formed that the Indians would	attack	and destroy the new settlement of buffalo hunters
by Isa-tai and famed Comanche chief Quanah Parker,	attacked	the Adobe Walls post. The Indians planned to catch
U.S. Army for cavalry use on the Plains. The	attack	on Adobe Walls by the Indians served as a catalyst
Indian prophet Isa-tai, he led some 300 Indians in an	attack	on buffalo hunters in the Adobe Walls post. <p
tan and his angels, they will surround the Holy City to	attack	God and His people. God will destroy Satan and his
aming. A lot of VCs (venture capitalists) are probably	attacking	this theme. Facebook gaming is massive - but mol
n organised group or government that commits cyber	attacks	], and sadly 2012 will be the year that advanced pe
m Fix, a victim of an acute myocardial infarct (heart	attack	). Closer to home, a cardiologist at a local hospital
s devotion to regular exercise, also died from a heart	attack	. Clearly, lack of exercise was not a factor in thes
nto them. The majority of computer viruses	attack	computers through e-mail attachments, so you sh
ore sophisticated - some viruses are even designed to	attack	Outlook address books, which means that the virus
> Firstly, I am extremely disappointed at the personal	attack	you have made on my integrity and to what I can
l leaders, that when the argument resorts to personal	attack	, the argument is lost by the detractor. How can a
rail: dry / snow free * It seems like there is a pack of	attack	/ guard grouse on the trail. They are pretty aggre
e. Do they make grouse spray? Alpine/ Peak:	Attack	grouse on the Col. Lots of deep snow towards the
smoker, you're five times more likely to have a heart	attack	than a non-smoker of the same age. Smo
rance will improve. After one year, your risk of heart	attack	is cut in half compared with that of a smoker. An
that the town wouldn't refurbish Engine 2 - the oldest	attack	engine in the department's fleet. Selectm
en written by concerned parents, and after the initial	attack	on morals, judgement and taste, most correspond
ic treatment is given via the bloodstream or orally to	attack	the cancerous cells within the body and takes the
tectomy or radical mastectomy. All of these surgeries	attack	the lump or mass directly and the amount of mass
the Vaughan Road from his unsuccessful raid and was	attacked	by elements of Brig. Gen. John Pegram's Confeder

attack *(verb)*

English Web 2013 (enTenTen13) freq = 883,561 (38.87 per million)

modifiers of "attack"			objects of "attack"			subjects of "attack"			"attack" and/or ...			prepositional phrases		
	128,071	14.49		490,583	55.52		222,347	25.16		51,372	5.81		109,110	
not +	20,050	2.09	people +	7,629	4.85	man +	3,813	4.90	kill +	3,937	9.09	"attack" by ...	56,428	6.39
not attack			problem +	4,531	4.89	people +	3,418	2.92	attacked and killed			"attack" in ...	11,647	1.32
also +	5,934	2.47	enemy +	4,363	7.48	force +	3,310	7.09	destroy +	2,247	8.92	"attack" with ...	9,629	1.09
also attacked			iran +	4,183	7.93	forces attacked			attack and destroy			"attack" on ...	6,772	0.77
then +	5,087	3.97	to attack Iran			dog +	3,036	6.63	defend +	2,246	9.45	"attack" from ...	6,092	0.69
then attacked			force +	3,854	6.52	group +	2,444	5.04	threaten +	1,025	8.46	"attack" at ...	3,845	0.44
again +	4,234	5.34	attacking force			attacked by a group of			beat +	923	8.20	"attack" for ...	2,082	0.24
attacked again			system +	3,774	4.79	terrorist +	2,395	8.02	Is attacked and beaten			"attack" as ...	1,634	0.18
only +	3,301	2.88	man +	3,770	5.07	terrorists attacked			bite +	723	8.31	"attack" to ...	1,083	0.12
only attack			target +	3,736	6.98	enemy +	1,905	7.27	try +	589	4.74	"attack" during ...	847	0.10
just +	2,645	1.89	cell +	3,578	6.82	system +	1,896	4.48	to try and attack			"attack" without ...	844	0.10
just attack			city +	3,327	6.12	immune system attacks			injure +	581	7.61	"attack" through ...	745	0.08
physically +	2,477	7.31	attacking the city			army +	1,503	6.65	attacked and injured			"attack" of ...	708	0.08
physically attacked			israel +	3,247	7.34	mob +	1,437	7.52	move +	525	6.05	"attack" over ...	452	0.05
never +	2,247	2.78	to attack Israel			someone +	1,425	4.09	move and attack			"attack" after ...	427	0.05
never attacked			player +	3,128	5.80	virus +	1,417	6.99	rob +	512	7.99	"attack" before ...	380	0.04
directly +	2,224	5.43	attacking players			virus attacks			attacked and robbed			"attack" if ...	331	0.04
directly attack			woman +	3,074	5.21	police +	1,393	5.58	burn +	486	7.02	"attack" than ...	330	0.04
viciously +	2,089	8.90	ship +	2,912	6.77	police attacked			attacked and burned			"attack" like ...	319	0.04
viciously attacked			country +	2,753	5.65	bomber +	1,298	7.39	harass +	484	7.71	"attack" while ...	319	0.04
even +	2,040	2.67	person +	2,705	5.05	bombers attacked			capture +	450	6.81	"attack" because ...	311	0.04
even attack			village +	2,427	6.74	militant +	1,213	7.27	attacked and captured			"attack" along ...	272	0.03
often +	1,781	3.05	anyone +	2,284	5.51	militants attacked			take +	446	5.16	"attack" across ...	265	0.03
often attacked			attack anyone			israel +	1,035	5.97	rape +	438	7.48	"attack" into ...	254	0.03

Word list

Corpus: **English Web 2013 (enTenTen13)**
Total number of items: **6,577,602**

Page 1 Go Next >

word	frequency
the	964,787,346
to	558,301,035
and	553,532,182
of	516,717,484
a	427,921,752
in	318,494,241
is	245,015,015
that	215,149,066
for	195,970,779
I	174,347,982
you	160,428,344
with	149,250,264
it	138,417,245
on	138,409,664
are	125,108,605
The	120,204,565
be	119,613,102
as	119,300,060
was	100,497,282
have	97,328,322
or	89,181,343
this	85,899,632
at	82,939,458
not	82,625,892
your	82,027,624
from	81,389,534
by	76,204,657
will	73,924,331
can	71,664,613
an	67,319,888
they	58,495,383

Word list

Corpus: **English Web 2013 (enTenTen13)**

Page 1 Go Next >

word (n-grams)	frequency
of the	10,736,346
in the	7,401,945
to the	4,561,848
on the	3,689,147
to be	3,111,878
for the	2,883,644
and the	2,711,147
is a	2,528,059
with the	2,271,059
at the	2,082,967
from the	1,887,184
in a	1,856,574
will be	1,697,207
of a	1,667,069
is the	1,650,926
it is	1,647,064
with a	1,568,943
for a	1,500,611
that the	1,486,036
as a	1,447,567
do n't	1,433,406
you can	1,407,847
can be	1,369,156
by the	1,357,868
you are	1,222,505
one of	1,194,358
If you	1,176,251
if you	1,107,220
to get	1,100,275
as well	1,094,881
to a	1,091,423
have a	1,071,729

attack *(verb)*

English Web 2013 (enTenTen13) freq = 883,561 (38.87 per million)

atacar *(verb)*

American Spanish Web 2011 (esamTenTen11, Freeling v4) freq = 403,298 (46.67 per million)

Use another candidate translation: ataque atentado terrorista contra septiembre atentar condenar terrorismo Israel

Click on collocates to access reciprocal bilingual search

modifiers of "attack"		
	128,071	14.49
viciously	2,089	8.90
viciously attacked		
brutally	1,606	8.11
brutally attacked		
verbally	1,297	7.94
verbally attacked		
violently	1,215	7.75
violently attacked		
physically	2,477	7.31
physically attacked		
savagely	584	7.12
savagely attacked		
aggressively	908	7.01
aggressively attack		
relentlessly	621	6.95
relentlessly attacked		
allegedly	953	6.53
allegedly attacked		
mistakenly	493	6.39
immune system mistakenly attacks		
repeatedly	1,164	6.32
repeatedly attacked		
personally	1,690	6.11
personally attacked		
mercilessly	298	6.08

objects of "attack"		
	490,583	55.52
iran	4,183	7.93
to attack Iran		
enemy	4,363	7.48
israel	3,247	7.34
to attack Israel		
iraq	2,229	7.05
to attack Iraq		
target	3,736	6.98
midfielder	1,900	6.95
an attacking midfielder		
cell	3,578	6.82
ship	2,912	6.77
syria	1,714	6.75
to attack Syria		
village	2,427	6.74
opponent	2,139	6.70
force	3,854	6.52
attacking force		
convoy	1,394	6.49
civilian	1,287	6.23
attacking civilians		
states	1,498	6.19
to attack the United States		
harbor	1,166	6.19
attacked Pearl Harbor		

object		
	82,749	20.52
causa	1,768	7.66
atacar las causas		
cuartel	480	7.43
raíz	606	7.37
atacar la raíz		
aro	442	7.34
atacar el aro		
problema	5,576	7.01
atacar el problema		
irán	313	6.90
atacar Irán		
campamento	375	6.88
pobreza	517	6.69
para atacar la pobreza		
célula	383	6.63
ataca las células		
convoy	258	6.63
sede	407	6.62
atacaron la sede		
blanco	296	6.58
atacar blancos		
corrupción	359	6.52
atacar la corrupción		
ciudad	858	6.49
embajada	246	6.46

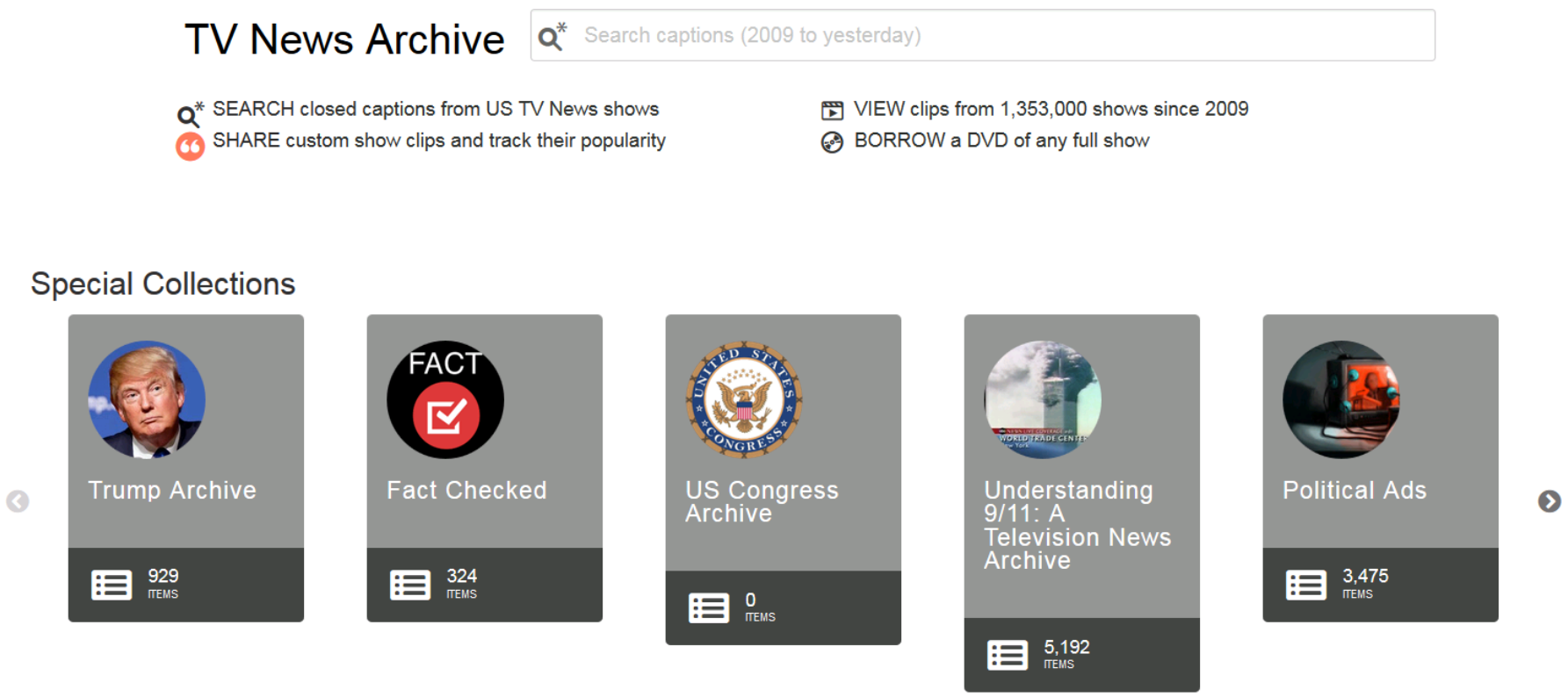
subjects of "attack"		
	222,347	25.16
terrorist	2,395	8.02
terrorists attacked		
mob	1,437	7.52
bomber	1,298	7.39
bombers attacked		
enemy	1,905	7.27
militant	1,213	7.27
militants attacked		
force	3,310	7.09
forces attacked		
gunman	1,033	7.09
gunmen attacked		
panic	1,007	7.04
panic attacks		
virus	1,417	6.99
virus attacks		
pirate	890	6.79
attacked by pirates		
shark	916	6.74
attacked by a shark		
army	1,503	6.65
dog	3,036	6.63
monster	945	6.61
hacker	842	6.61
thug	601	6.33

subject_np		
	31,082	7.71
virus	334	7.51
virus ataca		
otan	224	7.35
de la OTAN atacaron		
comando	226	7.34
un comando atacó		
farc	259	7.31
de las Farc atacaron		
perro	396	7.23
fuerza	669	7.13
enemigo	223	7.13
ejército	355	7.04
aviación	139	6.99
tropa	199	6.97
avión	201	6.76
rebelde	104	6.41
los rebeldes atacaron		
terrorista	101	6.38
los terroristas atacaron		
tiburón	86	6.35
grupo	1,083	6.34
pirata	87	6.27
insurgente	80	6.25
los insurgentes atacaron		
enfermedad	264	6.21

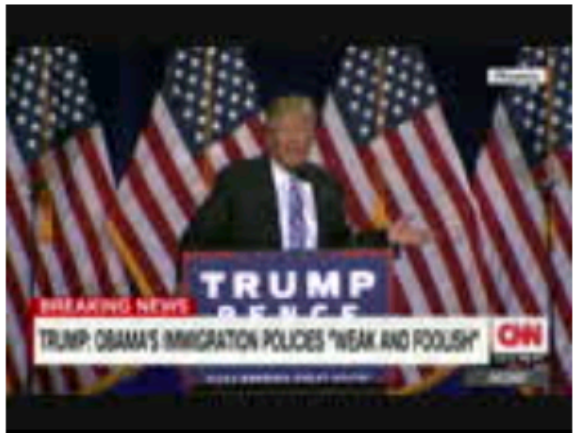
y_o		
	14,506	3.60
destruir	657	9.25
atacar y destruir		
defender	757	9.18
atacar y defender		
matar	373	8.57
incendiar	157	8.30
atacaron e incendiaron		
defender+se	150	8.27
atacar y defenderse		
herir	169	7.98
atacado y herido		
insultar	118	7.65
atacar e insultar		
asesinar	195	7.64
descalificar	106	7.64
atacar y descalificar		
amenazar	168	7.59
saquear	99	7.56
perseguir	122	7.51
invadir	92	7.51
atacar e invadir		
agredir	109	7.50
golpear	134	7.43
atacado y golpeado		
criticar	137	7.41

TV News Archive

https://archive.org/details/tv

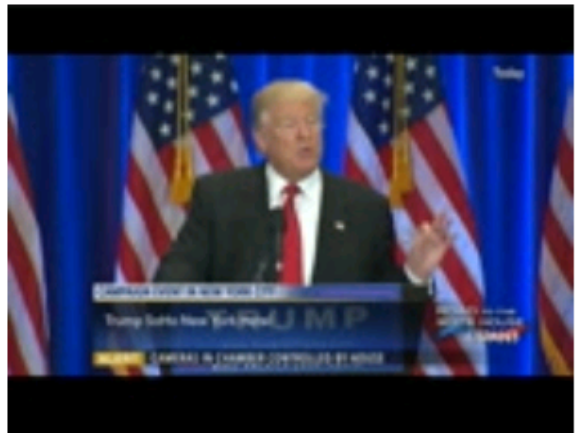


Search Results 0 to 49 of about 475 *(some duplicates have been removed)*



Anderson Cooper 360
CNN
Aug 31, 2016 10:00pm PDT

“ to their names. 25,000. on top of that, illegal immigration costs our country more than $113 billion a year.



Politics and Public Policy Today
CSPAN
Jun 22, 2016 4:01pm EDT

“ immigration security, and hillary wants neither. hillary clinton has put forward the most radical immigration platform in the history of the united states.

Key Capitol Hill Hearings
CSPAN
Oct 20, 2016 12:35pm EDT

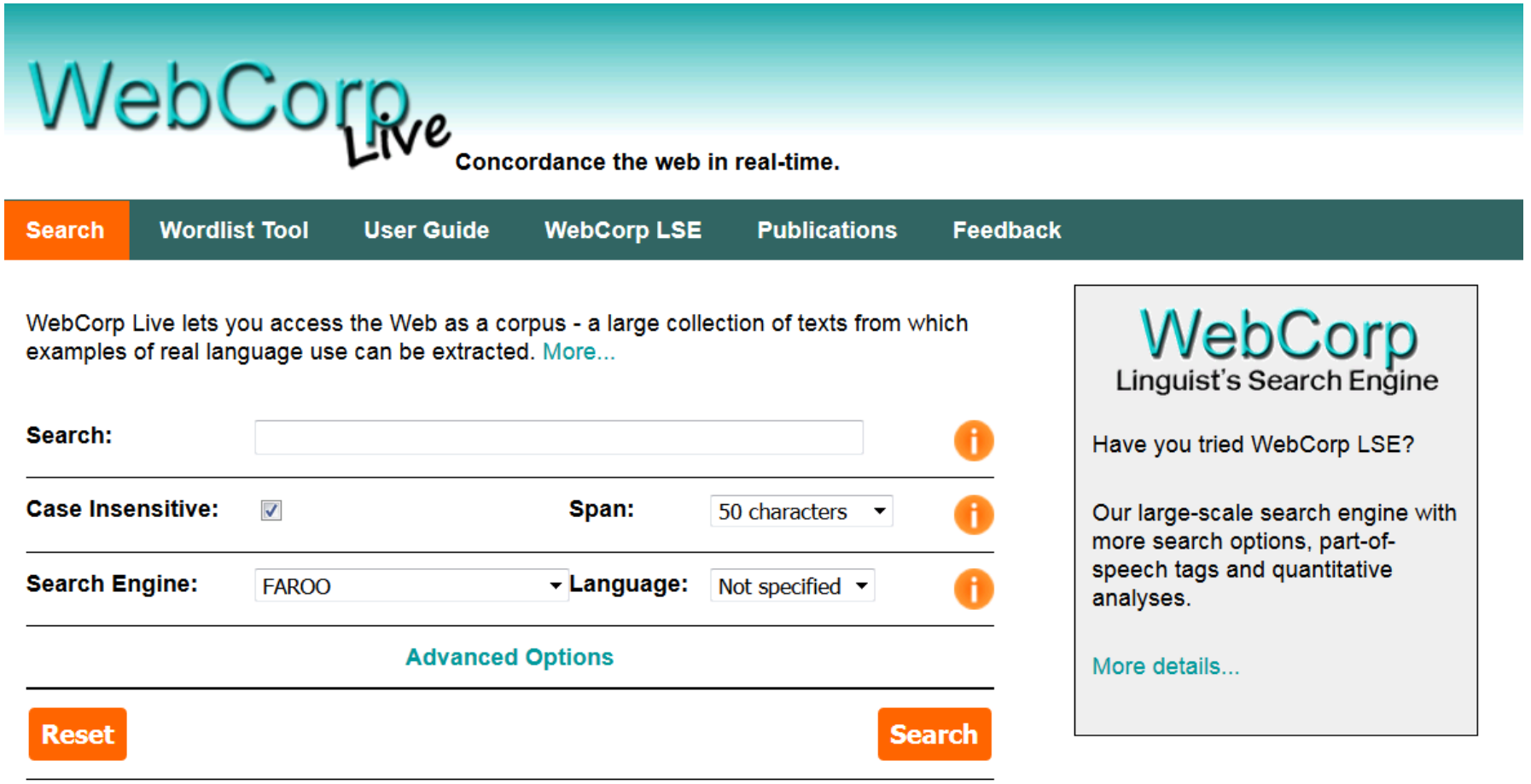
“ , was quoted in wikileaks as saying illegal immigrants could vote as long as

The Last Word With Lawrence O Donnell
MSNBC
Aug 31, 2016 7:01pm PDT

“ we will enforce all of our immigration laws. and the same goes for government benefits. the center for immigration studies estimates that 62% of households

Web-as-Corpus

http://www.webcorp.org.uk/live/

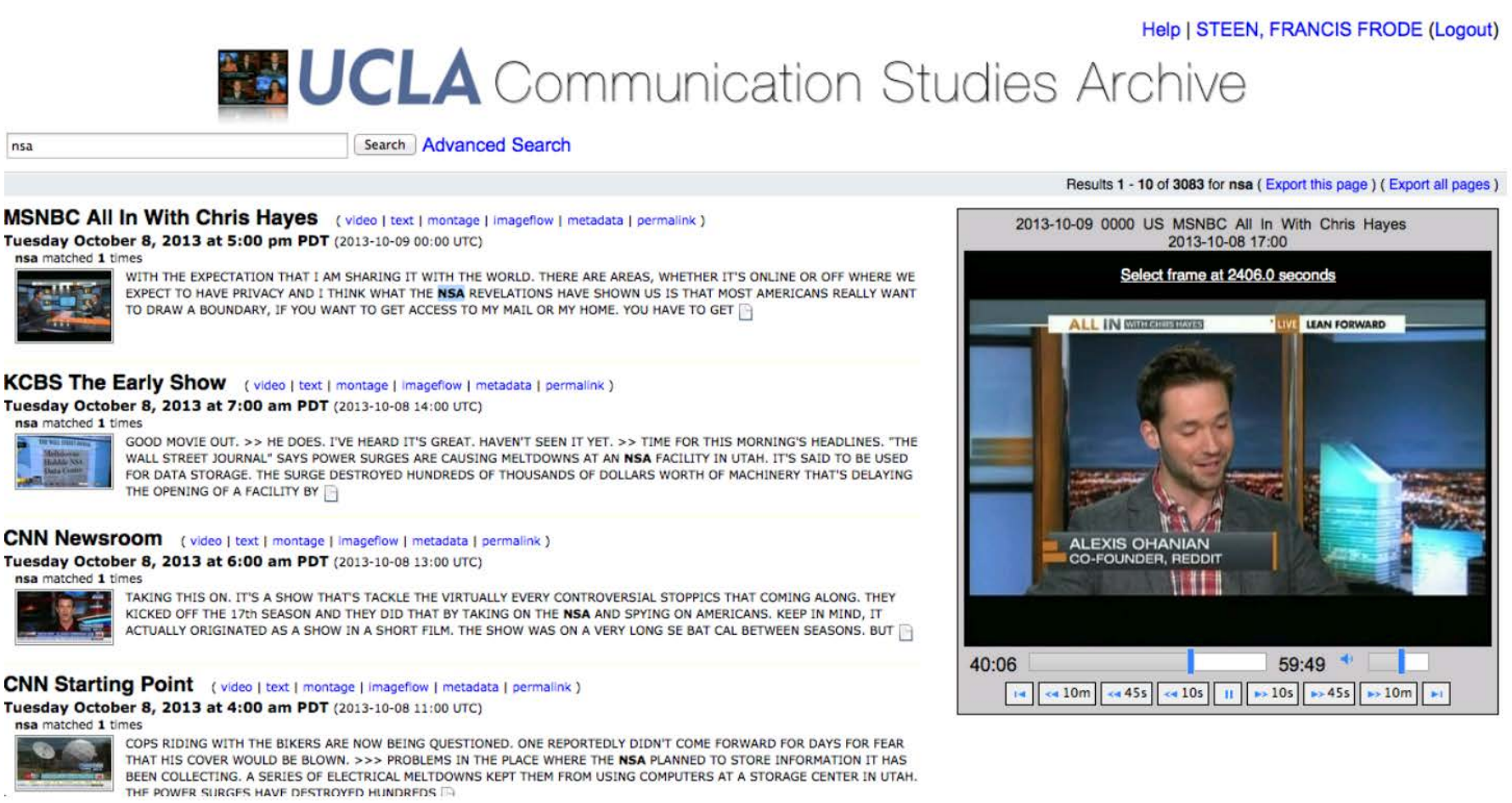


Other Corpus Software

WordSmith Tools	(GPB 50)	http://www.lexically.net/wordsmith/
AntConc	(free)	http://www.laurenceanthony.net/software/antconc/
MonoConc	($85)	http://www.athel.com/mono.html

Multimodal Analysis

Red Hen Lab http://www.redhenlab.org/ (Mark Turner, Case Western University)



Exploratory Corpus Linguistics

Exploratory – you do not have a particular research question in mind; you are just looking for anything cool.

Voyant tools: http://voyant-tools.org/

VOYANT

see through your text

https://www.nytimes.com/2017/04/04/us/politics/us-appetite-for-mexicos-drugs-fuels-illegal-immigration.html?rref=collection%2Fsectioncollection%2Fus&action=click&contentCollection=us&region=rank&module=package&version=highlights&contentPlacement=1&pgtype=sectionfront

Word cloud

Source text

Reader

LUKEVILLE, Ariz. — Each year, Mexican drug cart...

LUKEVILLE, Ariz. — Each year, Mexican drug cartels rake in billions of dollars in profits from the sale of heroin, methamphetamines and other drugs in the United States. The money has to make its way south somehow.

Though the cartels sometimes hire legitimate companies to buy goods like silk and ink cartridges and export them to Mexico, where they are sold for pesos, a more common method is to simply pay someone to drive the cash over the border.

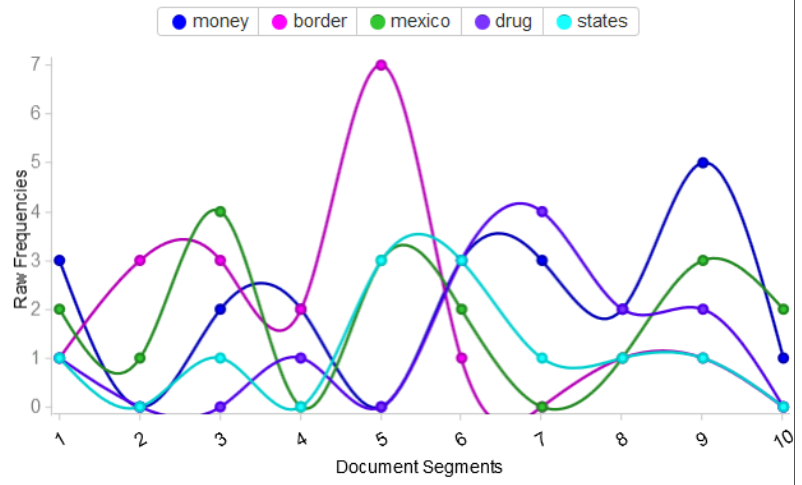
President Trump has talked frequently about “bad hombres” streaming in from Mexico. But it is the flow of money going from north to south — a product of Americans’ voracious appetite for illicit drugs — that officials say is an equal part of the problem.

“It’s the money and the guns that have enabled the cartels to obtain the power they have,” Scott Brown, special agent in charge of Homeland Security Investigations in Phoenix, part of Immigration and Customs Enforcement, said in an interview. “I’m a

Terms:

Segmented freq

Trends | Document Terms



Frequencies | Segments:

Vocab density and freq

Summary | Documents | Phrases

This corpus has 1 document with 1,616 total words and 683 unique word forms. Created 22 seconds ago.

Vocabulary Density: 0.423

Average Words Per Sentence: 25.7

Most frequent words in the corpus: money (21); border (19); mexico (18); drug (13); states (11)

Concordance

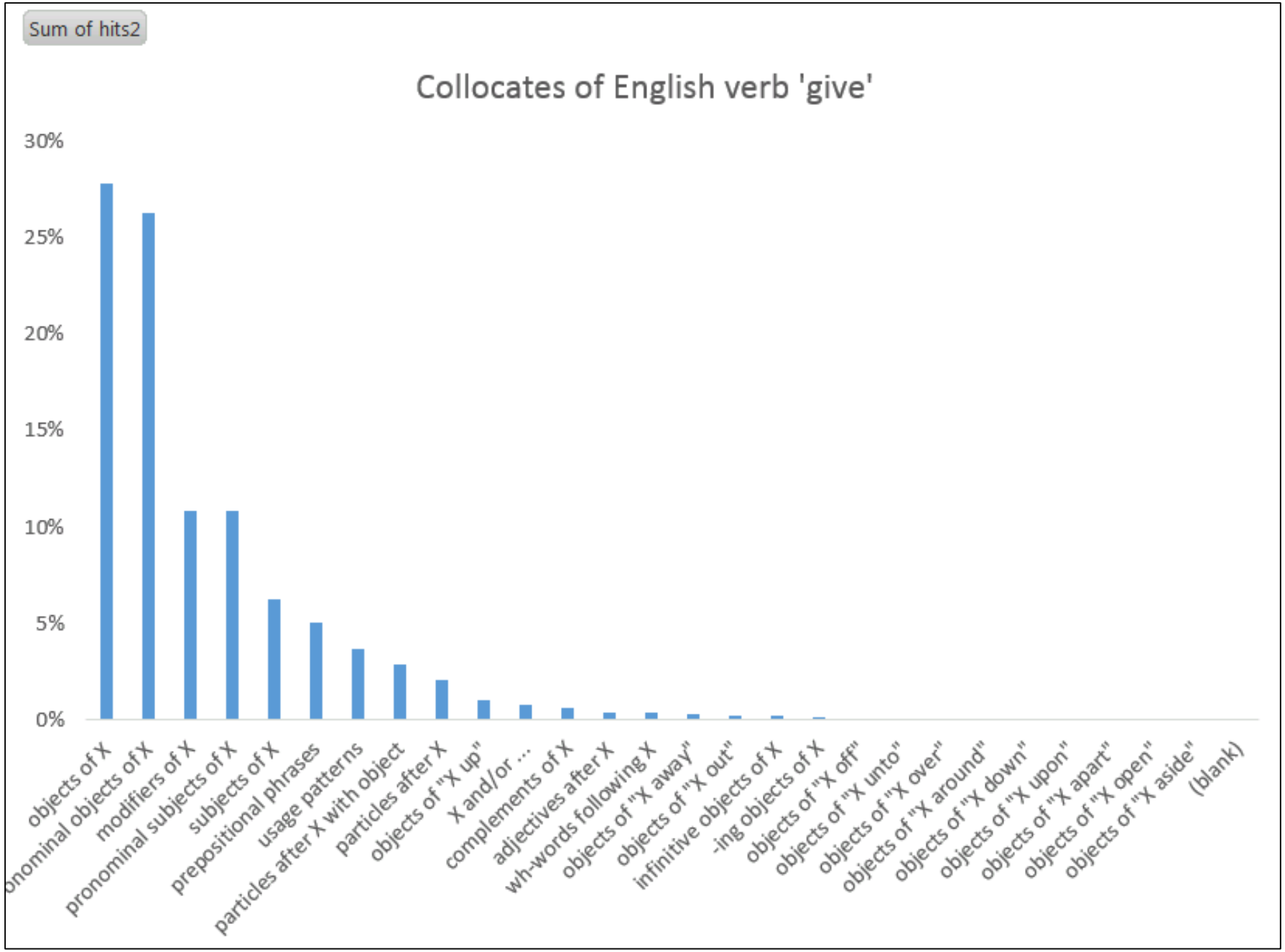
Contexts | Bubblelines

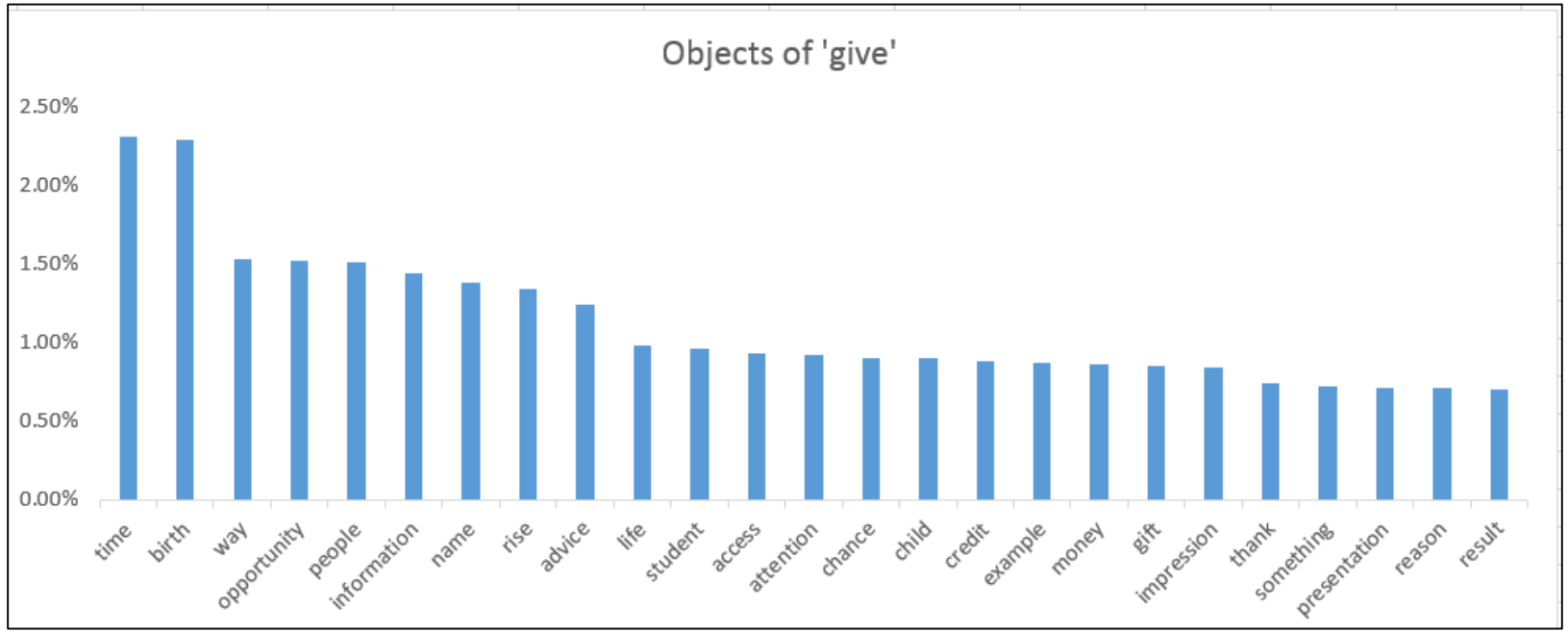
Document	Left	Term	Right
1) LUKE…	in the United States. The	mo…	has to make its way
1) LUKE…	it is the flow of	mo…	going from north to south
1) LUKE…	of the problem. “It’s the	mo…	and the guns that have
1) LUKE…	going south for guns and	mo…	,” Pete Bachelier, a United States
1) LUKE…	the flow of guns and	mo…	entering the country from the
1) LUKE…	border patrol agents say that	mo…	, which was found while following
1) LUKE…	more officers to look for	mo…	and guns amid a decline

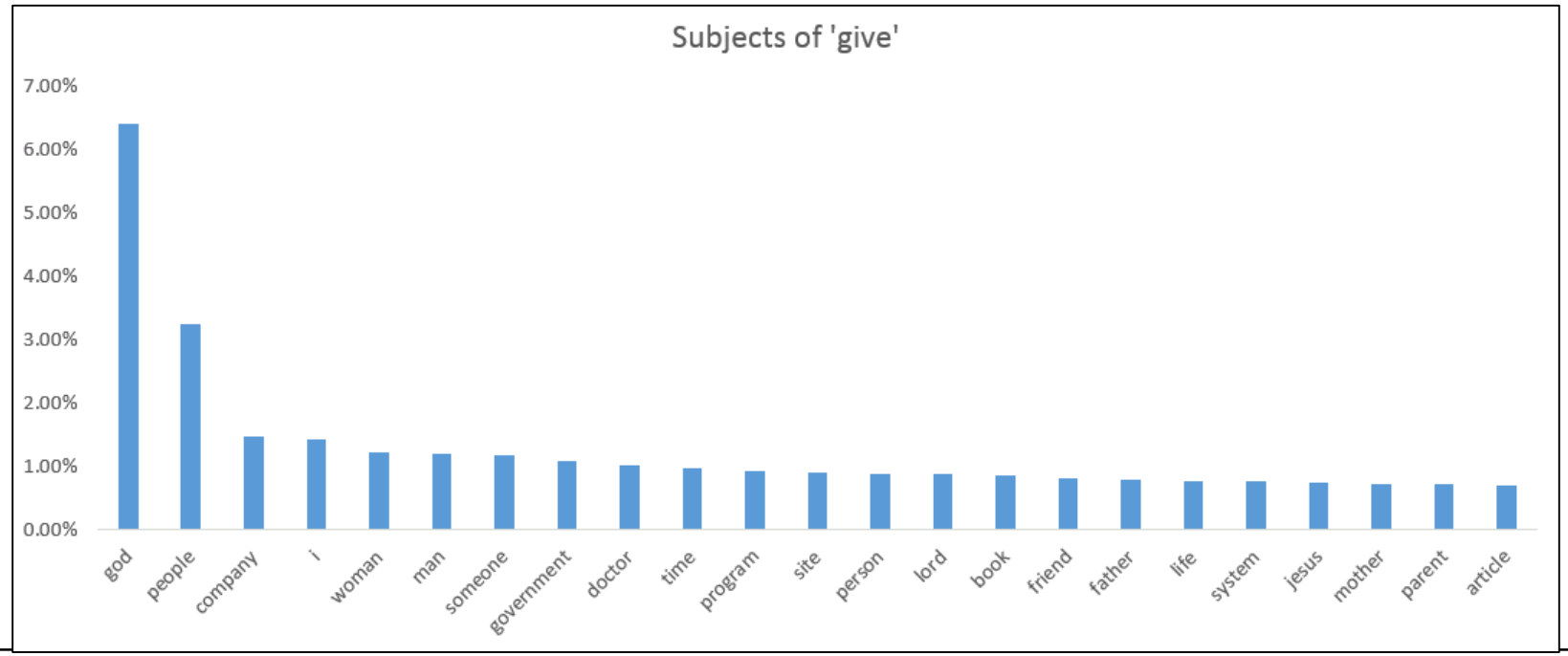
Organizing and Processing The data

Organizing and Processing

- General WYSIWYG – store files locally, process through generally-available programs like Excel.
- Specialized WYSIWYG – work directly in the corpus tool, e.g., in WordSmith Tools or in AntConc.
- Programming – Use R and Python


Sum of hits2
Collocates of English verb 'give'
30%
25%
20%
15%
10%
5%
0%
objects of X
pronominal objects of X
modifiers of X
pronominal subjects of X
subjects of X
prepositional phrases
usage patterns
particles after X with object
particles after X
objects of "X up"
X and/or ...
complements of X
adjectives after X
wh-words following X
objects of "X away"
objects of "X out"
infinitive objects of X
-ing objects of X
objects of "X off"
objects of "X unto"
objects of "X over"
objects of "X around"
objects of "X down"
objects of "X upon"
objects of "X apart"
objects of "X open"
objects of "X aside"
(blank)


Objects of 'give'
2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
time
birth
way
opportunity
people
information
name
rise
advice
life
student
access
attention
chance
child
credit
example
money
gift
impression
thank
something
presentation
reason
result


Subjects of 'give'
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
god
people
company
i
woman
man
someone
government
doctor
time
program
site
person
lord
book
friend
father
life
system
jesus
mother
parent
article

Discourse Analysis

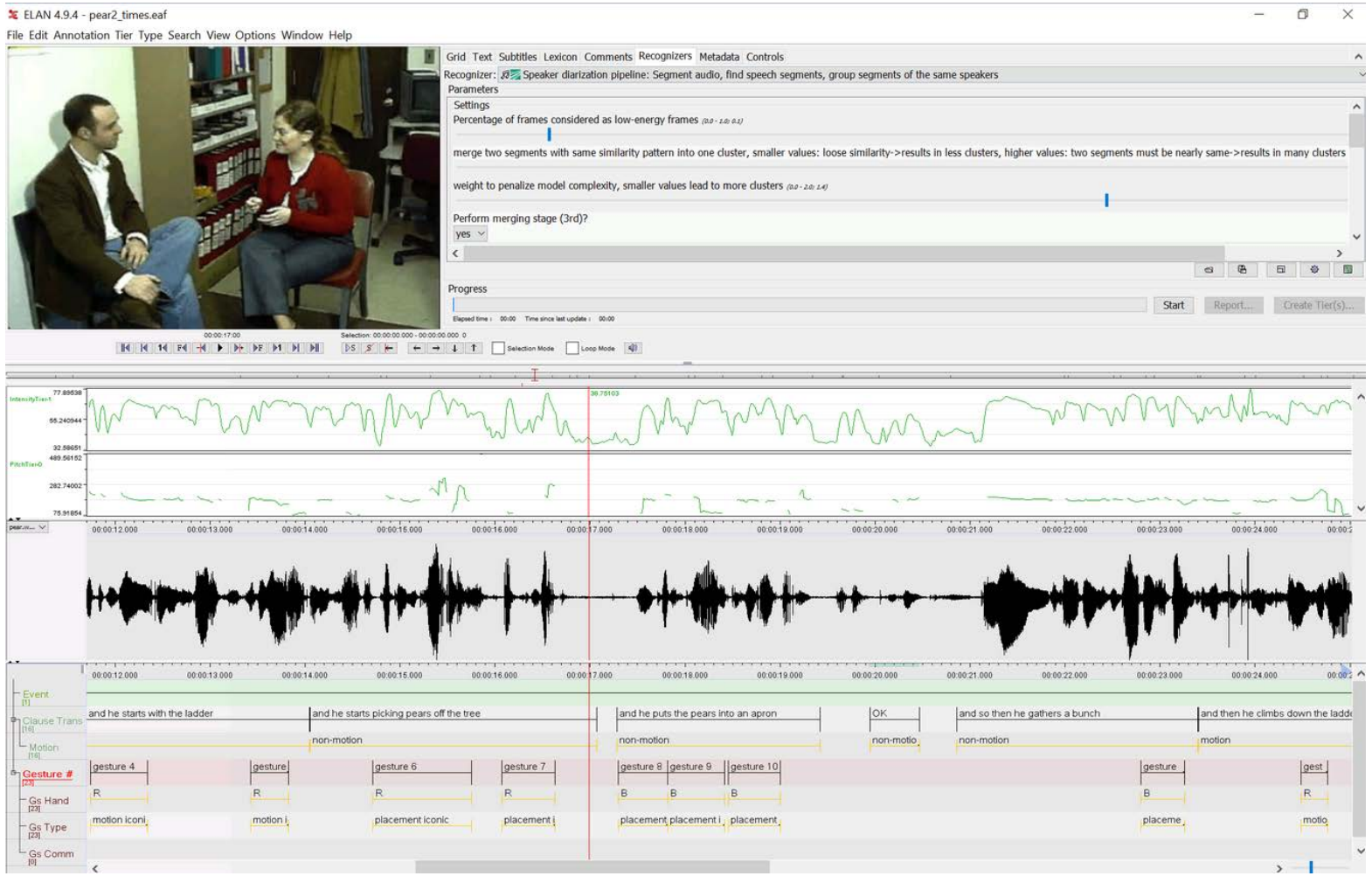
- The spoken word – already transcribed
  - Qualitative Data/Discourse Analysis (QDA) Tools
    - MaxQDA, Nvivo, QDA Miner (paid)
    - LibreQDA, ELAN, Free QDA

- The spoken word – not already transcribed
  - You need to transcribe
    - What transcription conventions?
    - Representing turn taking, interruptions, silence
    - Technical difficulties

Discourse and Multimodal Analysis

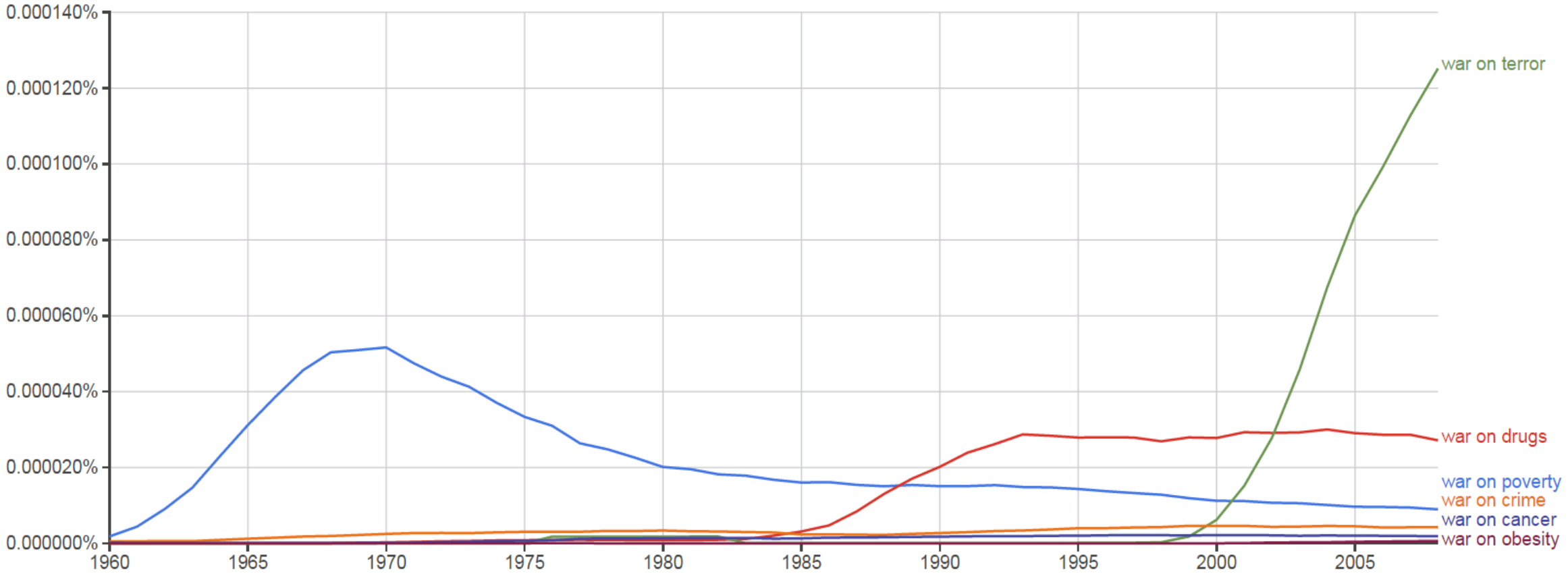
ELAN

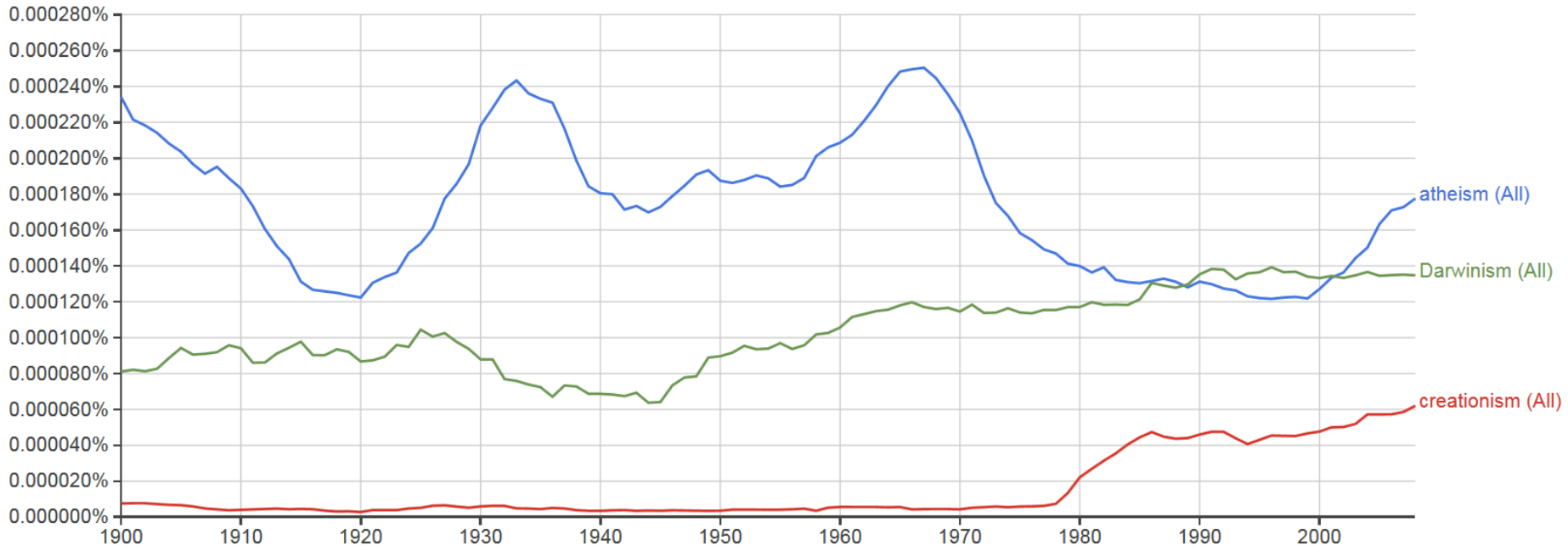
https://tla.mpi.nl/tools/tla-tools/elan/

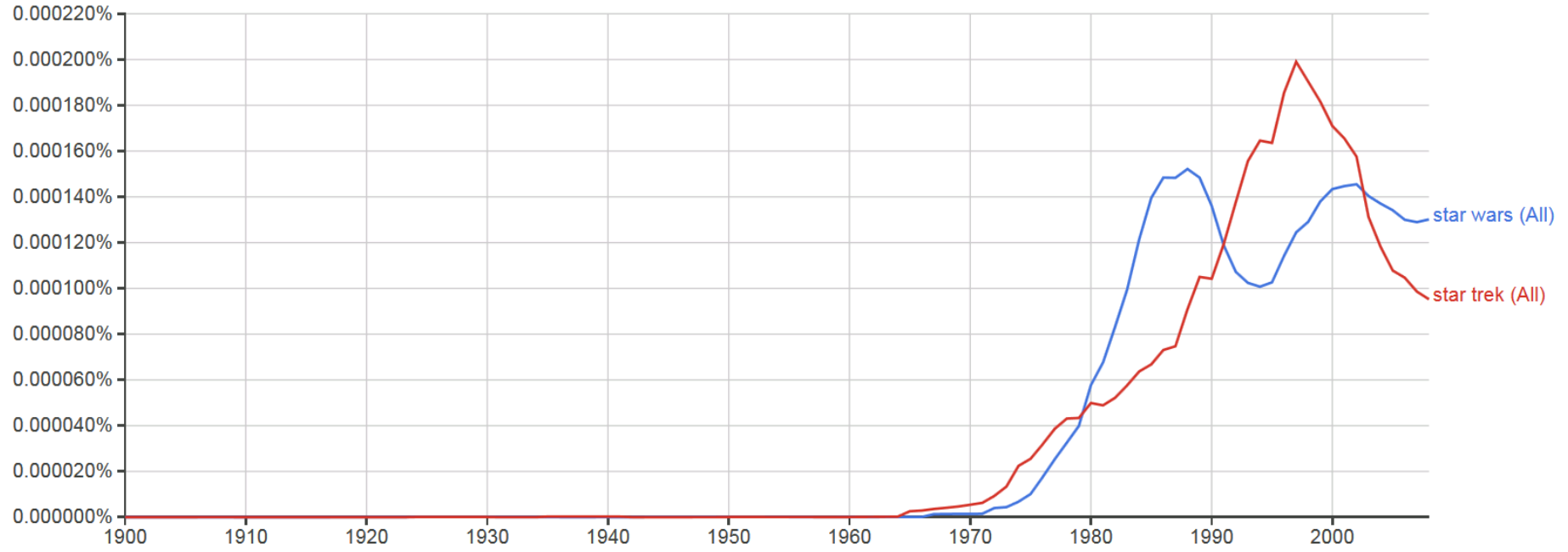


N-Grams

- Bigrams: two words, trigrams: three words, etc.




0.000280%
0.000260%
0.000240%
0.000220%
0.000200%
0.000180%
0.000160%
0.000140%
0.000120%
0.000100%
0.000080%
0.000060%
0.000040%
0.000020%
0.000000%
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
atheism (All)
Darwinism (All)
creationism (All)


0.000220%
0.000200%
0.000180%
0.000160%
0.000140%
0.000120%
0.000100%
0.000080%
0.000060%
0.000040%
0.000020%
0.000000%
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
star wars (All)
star trek (All)

Semantic Analysis

FrameNet	https://framenet.icsi.berkeley.edu/fndrupal/
MetaNet	https://metaphor.icsi.berkeley.edu/pub/en/
CAN	http://dev-constructional-annotation-net.pantheon.berkeley.edu/

Category:Metaphor

This is the Metaphor category page. This category uses the form Metaphor.

Create or edit a Metaphor:

Create or edit

Total number of Metaphor pages: 685

A

- ABILITY TO EVALUATE GOVERNMENT IS ABILITY TO SEE
- ABILITY TO EVALUATE IS ABILITY TO SEE
- ABILITY TO KNOW IS ABILITY TO SEE
- ABUSIVE POLITICAL LEADERS ARE PHYSICAL BULLIES
- ACCEPTING IS SWALLOWING
- ACCESS TO EDUCATION IS ACCESS TO AN OBJECT
- ACCESS TO KNOWLEDGE IS ACCESS TO AN OBJECT
- ACHIEVING A PURPOSE IS ACQUIRING A DESIRED OBJECT
- ACHIEVING A PURPOSE IS REACHING A DESTINATION
- ACHIEVING GUN RIGHTS IS REACHING A DESTINATION
- ACHIEVING POWER IS MOVING UPWARDS
- ACQUIRING IDEAS IS EATING
- ACTION IS BEING IN A LOCATION
- ACTION IS MOTION
- ACTION IS MOTION ALONG A PATH
- ACTION IS SELF-PROPELLED MOTION ALONG A PATH

E cont.

- EMOTIONALLY MILD IS EXPERIENCING MILD TEMPERATURE
- EMOTIONS ARE OBJECTS
- EMOTIONS ARE SUBSTANCES
- ENABLE ABILITY TO UTILIZE RESOURCES IS ENABLE ABILITY TO ACCESS LOCATIONS
- ENACTING LEGISLATION IS CAUSING MOTION ALONG A PATH
- END OF AN ACTION IS THE END OF A PATH
- ENJOYING A RIGHT IS POSSESSING AN OBJECT
- ENTITLEMENT HOLDERS ARE LEECHES
- ENVIRONMENTAL HARM IS PHYSICAL INJURY
- EQUALITY IS BEING ON THE SAME VERTICAL LEVEL
- ESSENTIAL IS INTERNAL
- EVENTS ARE OBJECTS
- EVIL IS DARK

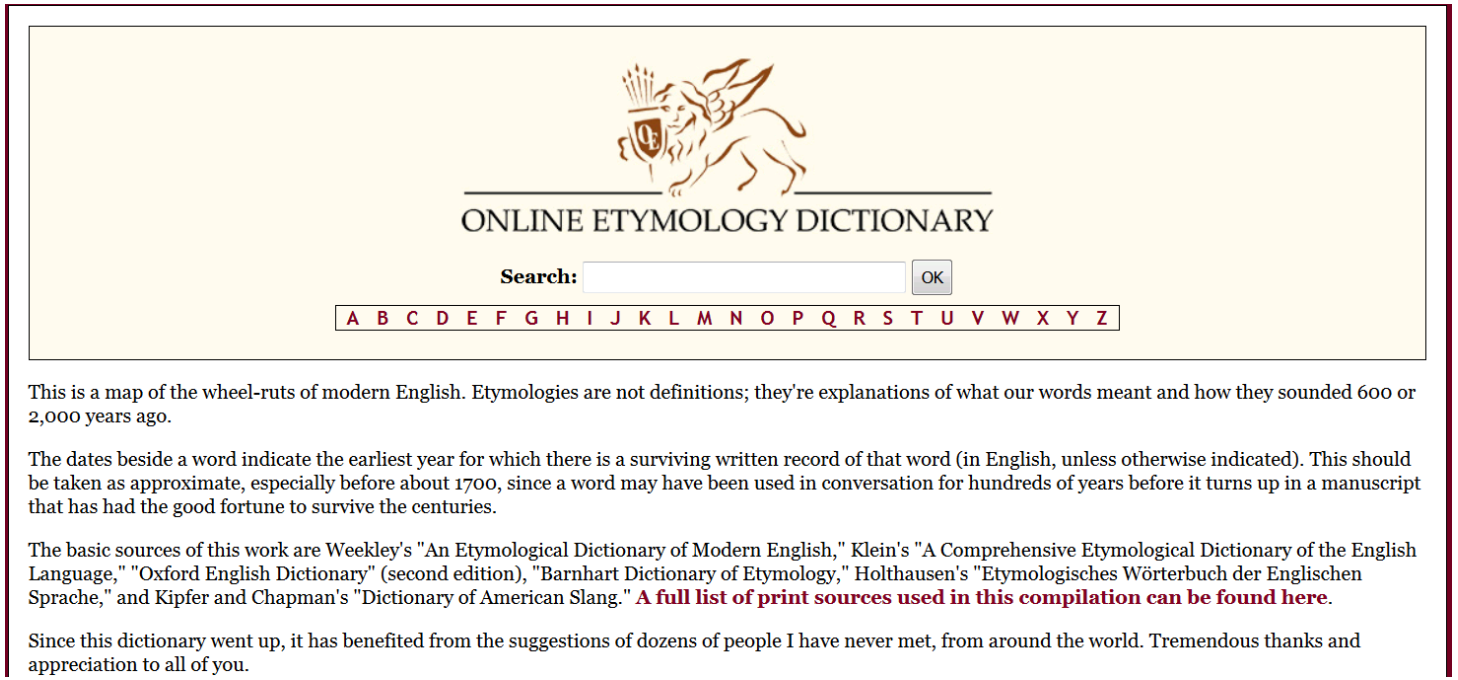
M

- MACHINES ARE PEOPLE
- MAINTAINING HIGHER SOCIAL STATUS IS MAINTAINING HOLD ON PHYSICAL POSITION
- MAKING A DECISION IS WEIGHING
- MAKING EDUCATIONAL OPPORTUNITY THE SAME FOR EVERYONE IS EQUALIZING QUANTITIES
- MAKING RELATED IS CAUSING CONTACT
- MANNER OF ACTION IS MANNER OF MOVEMENT
- MARKET IS A FAMILY
- MARKET IS A NATURAL PROCESS
- MARKET IS AN ACTOR
- MARRIAGE IS CLOTH
- MEANS ARE PATHS
- MEANS OF IMPROVING ECONOMIC STATUS ARE PATHS TO MOVING OUT OF A LOCATION
- MEMORIZATION IS WRITING

https://metaphor.icsi.berkeley.edu/pub/en/index.php/Category:Metaphor

Online Etymological Dictionary

http://www.etymonline.com/

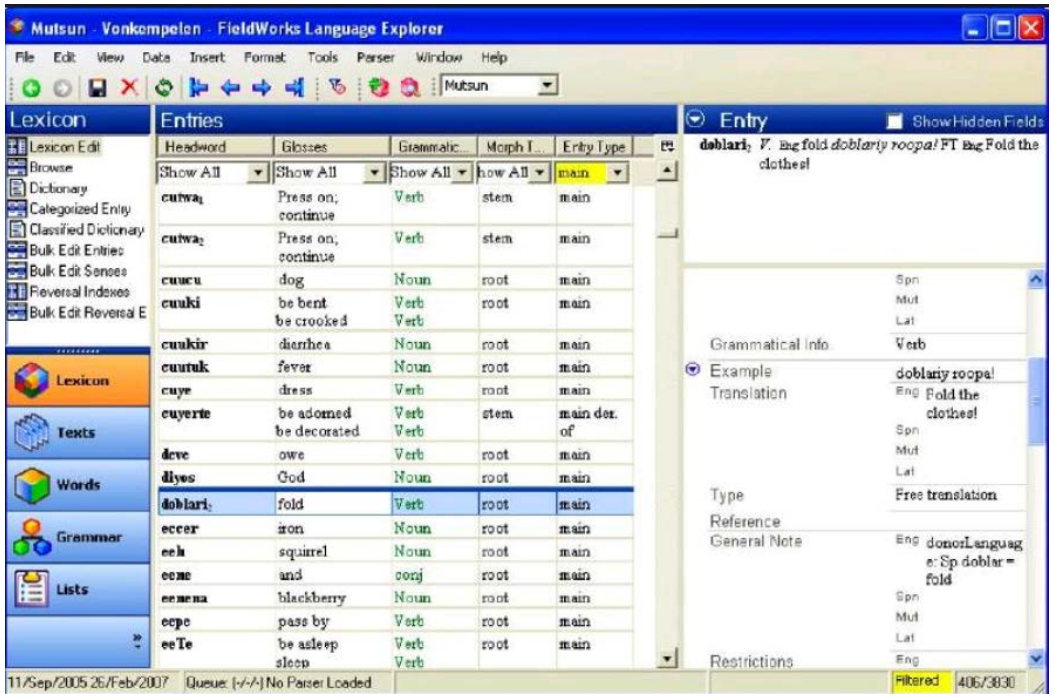


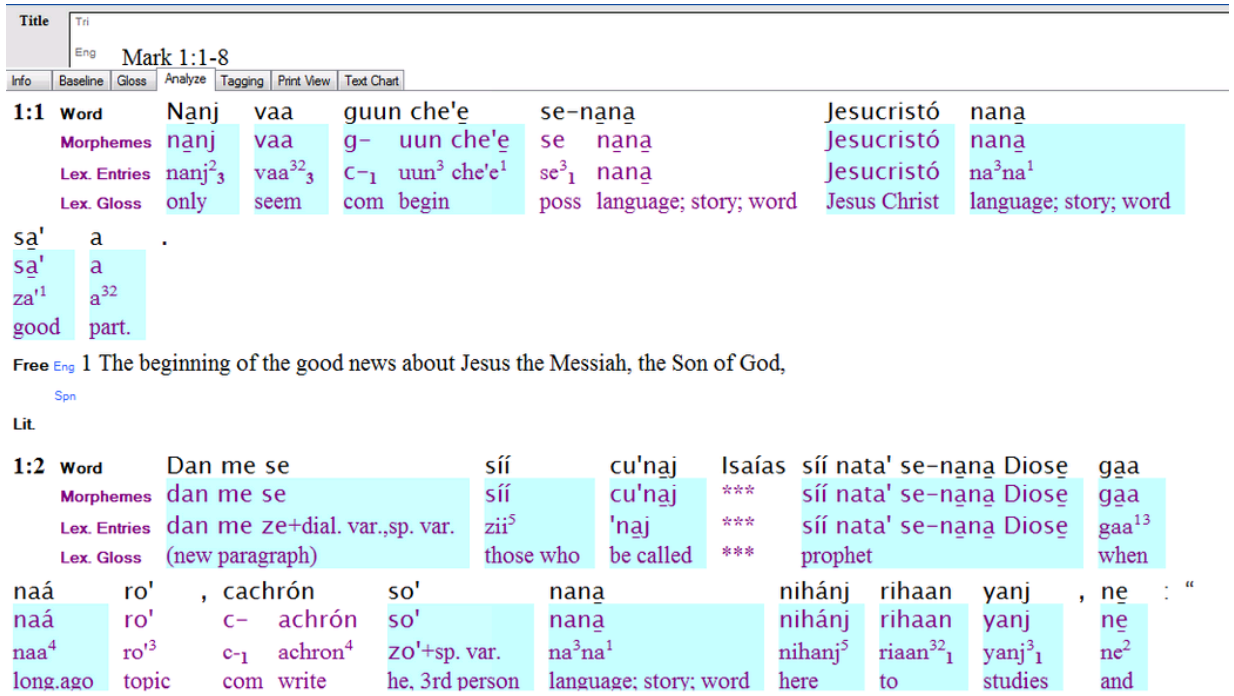
Multi-lingual Dictionary

http://www.wordreference.com/

Lexicons and Grammars

Flex (SIL FieldWorks) - http://fieldworks.sil.org/





If you want to work with linguistic data…

1. Figure out your **research question** first (but be flexible and allow for exploration).

2. Ask yourself, what kind of research question is this (what kind of data)?
   - Discourse
   - Word frequencies
   - Semantic breadth
   - Qualitative? Quantitative? Both?

3. You will work with a combination of tools; no single tool will do it all for you.